

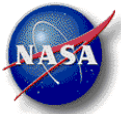
iRODS-Based Climate Data Services

John L. Schnase

*Office of Computational and Information
Science and Technology (Code 606)*

NASA Goddard Space Flight Center

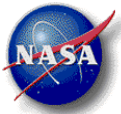
September 12, 2012



Climate Data and Climate Data Services



Edmund Halley's Wind Map (1686)
First global data image - a weather map, showing prevailing winds on a geographical map of the Earth.



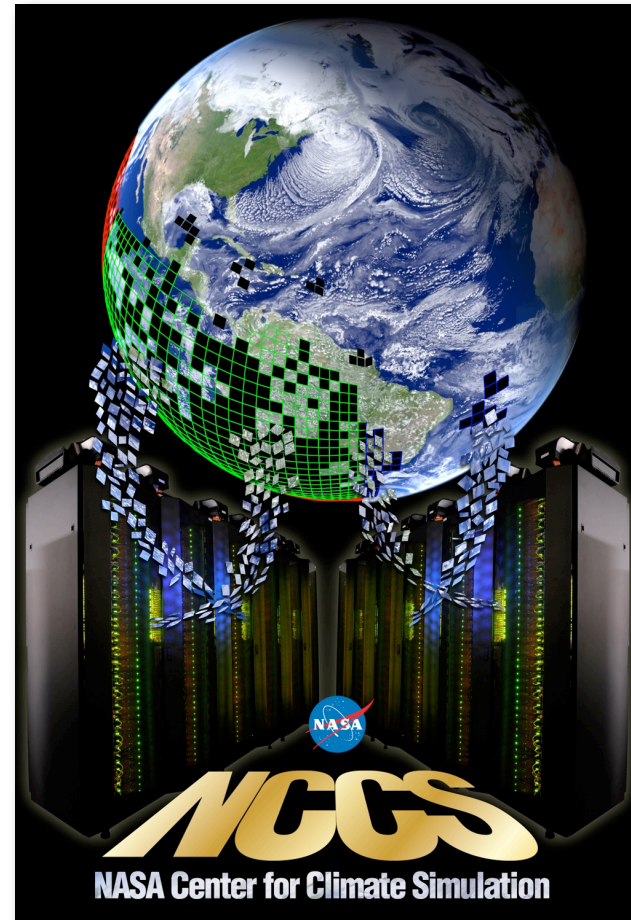
Climate Data and Climate Data Services

The story of climate data

- It's old, global in scale, and all of it is relevant.
- It's one of the fastest growing classes of scientific data.
- Climate research is becoming increasingly data centric.
- Data publication and data sharing is becoming a big deal.
- The use of climate data in other domains is expanding.

Data services as a core mission

- Climate data services represent an effort to respond to the story of climate data.
- The concept is being defined.
- At the very least, we know that it will require:
 - a new view of data stewardship,
 - new organizational processes, and
 - new data technologies.





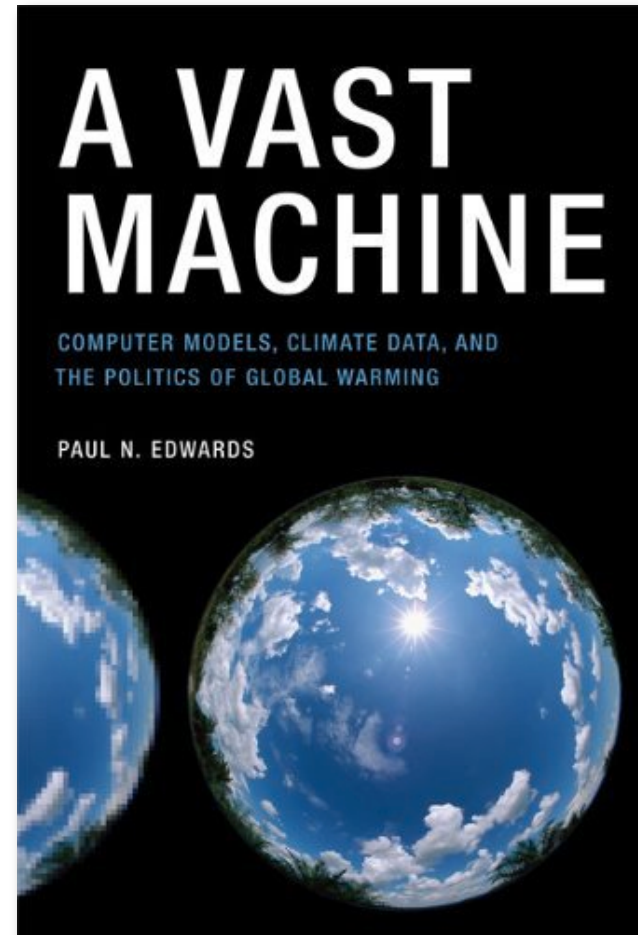
Climate Data and Climate Data Services

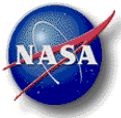
The story of climate data

- It's old, global in scale, and all of it is relevant.
- It's one of the fastest growing classes of scientific data.
- Climate research is becoming increasingly data centric.
- Data publication and data sharing is becoming a big deal.
- The use of climate data in other domains is expanding.

Data services as a core mission

- Climate data services represent an effort to respond to the story of climate data.
- The concept is being defined.
- At the very least, we know that it will require:
 - a new view of data stewardship,
 - new organizational processes, and
 - new data technologies.





Canonical Scenario

Scenario

A customer approaches the NCCS with a new dataset they want us to manage ...

Q. What technology is needed to quickly meet that customer's requirement under the follow constraints:

- The solution should be: simple, fast, and cheap;
- provide core capabilities to get started, but extendable to accommodate future needs;
- be flexible, with the ability to use, optimize, and change deployment configurations in response to resource availability; and
- allow the new dataset to be integrated into an existing data collection?

We're looking at solutions that combine data grids and cloud computing ...

Definitions

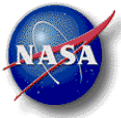
Customer – an individual scientist, a lab, project, or mission.

Dataset – may be products generated by a climate model, may be observational data, reanalysis data, or specialized products.

Manage – may refer to short-term file storage, long-term archival preservation; data may be used online by a person or application.

Examples

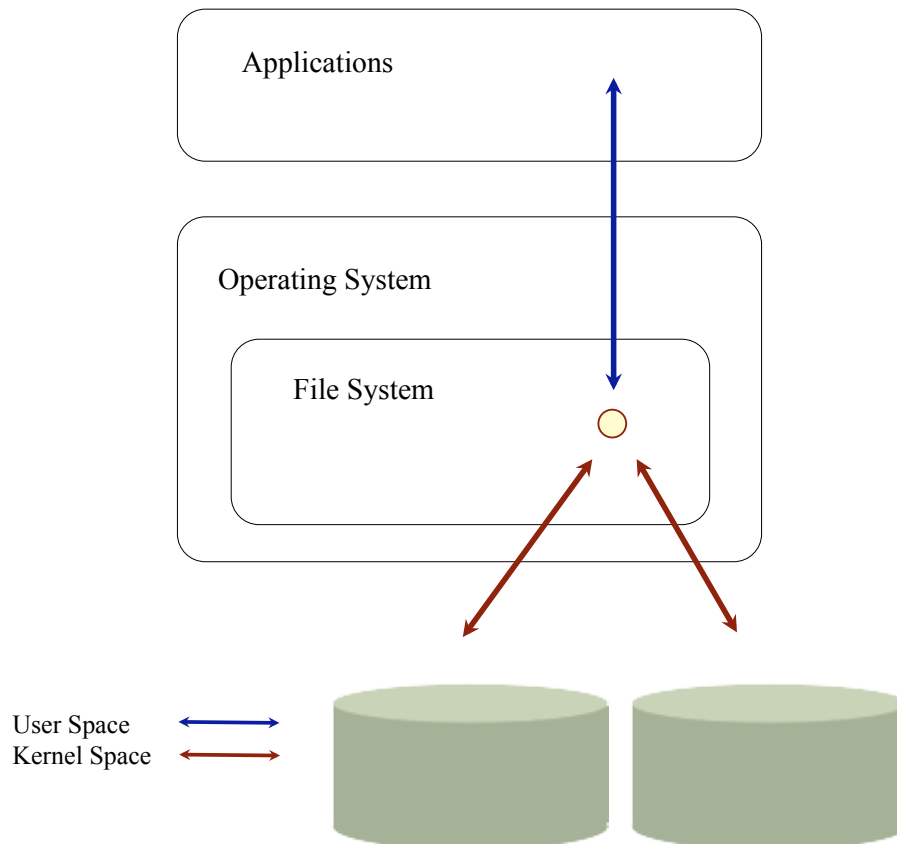
- IPCC AR5 data for ESGF.
- MODIS Atmospheres data for CMIP.
- AgMIP, SMOS, SMAP, TRMM, CERES, ...



Data Grid Software

POSIX Filesystem

A set of standard interfaces to the Unix operating system. Traditional filesystems are typically “owned” by the system ...

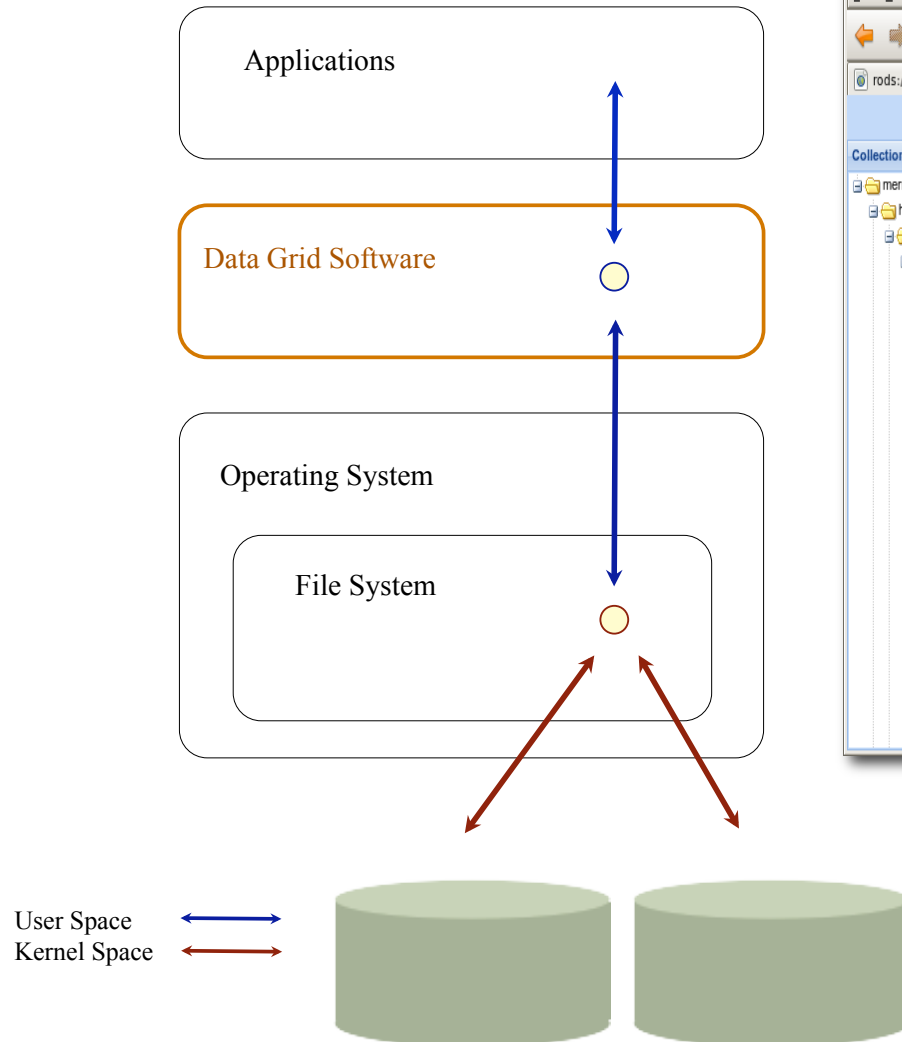


```
iShell
gs6060isfsn101:~ jschnase$ ls -al
total 440
drwxr-xr-x+ 33 jschnase staff 1122 Jul  8 12:28 .
drwxr-xr-x  9 root    admin  306 Feb 22 15:37 ..
-rw-r--r--  1 jschnase staff   3 Mar 24  2008 .CFUserTextEncoding
-rw-r--r--@  1 jschnase staff 12292 May 13 14:07 .DS_Store
-rw-r--r--  1 jschnase staff   0 Mar 24  2009 .Rhistory
drwx----- 2 jschnase staff   68 Jul  8 12:28 .Trash
drwxr-xr-x  2 jschnase staff   68 Apr 21  2008 .Xcode
-rw-r--r--  1 jschnase staff  5239 Jun 13 16:15 .bash_history
-rw-r--r--@  1 jschnase staff   58 Mar 23  2009 .bash_profile
-rw-r--r--@  1 jschnase staff   16 Mar 23  2009 .bashrc
drwx-----@ 3 jschnase staff  102 Mar 24  2008 .cups
drwx----- 10 jschnase staff   340 Jul  8 12:31 .dropbox
-rwxr-xr-x@  1 jschnase staff 10856 Mar 23  2009 .isfsrc
-rw-r--r--  1 jschnase staff   35 Nov 29  2010 .lessht
drwxr-xr-x  3 jschnase staff  102 Apr 28  2010 .agis
-rwxr-xr-x@  1 jschnase staff   15 Mar 23  2009 .rshrc
drwx-----@ 7 jschnase staff   238 Nov 29  2010 .ssh
-rw-r--r--  1 jschnase staff  2242 Nov 29  2010 .viminfo
drwxrwxrwx  3 jschnase staff   102 Mar 27  2008 Backups
drwx-----+ 5 jschnase staff   170 Jul  8 12:36 Desktop
drwx-----+ 17 jschnase staff   578 May 26 13:10 Documents
drwx-----+ 76 jschnase staff  2584 Apr 18 10:48 Downloads
drwx-----@ 18 jschnase staff   612 Jul  8 12:29 Dropbox
drwx-----+ 52 jschnase staff  1768 Mar  8 10:17 Library
drwx-----+ 3 jschnase staff   102 Mar 24  2008 Movies
drwx-----+ 4 jschnase staff   136 Apr  9  2008 Music
drwxr-xr-x 26 jschnase staff   884 Jun 13 13:31 NASA
drwxrwxrwx  9 jschnase staff   306 Mar 22 13:41 Nebula
drwxr-xr-x  5 jschnase staff   170 May  6  2010 Personal
drwx-----+ 18 jschnase staff   612 Jun 27 11:23 Pictures
drwxr-xr-x+  5 jschnase staff   170 Mar 24  2008 Public
drwxr-xr-x+ 16 jschnase staff   544 Jul 21  2008 Sites
drwxr-xr-x  4 jschnase staff   136 Aug 26  2010 _Desktop Archive
gs6060isfsn101:~ jschnase$
```

Standard POSIX filesystem ops and metadata over a traditional hierarchical filesystem ...

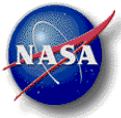


Data Grid Software

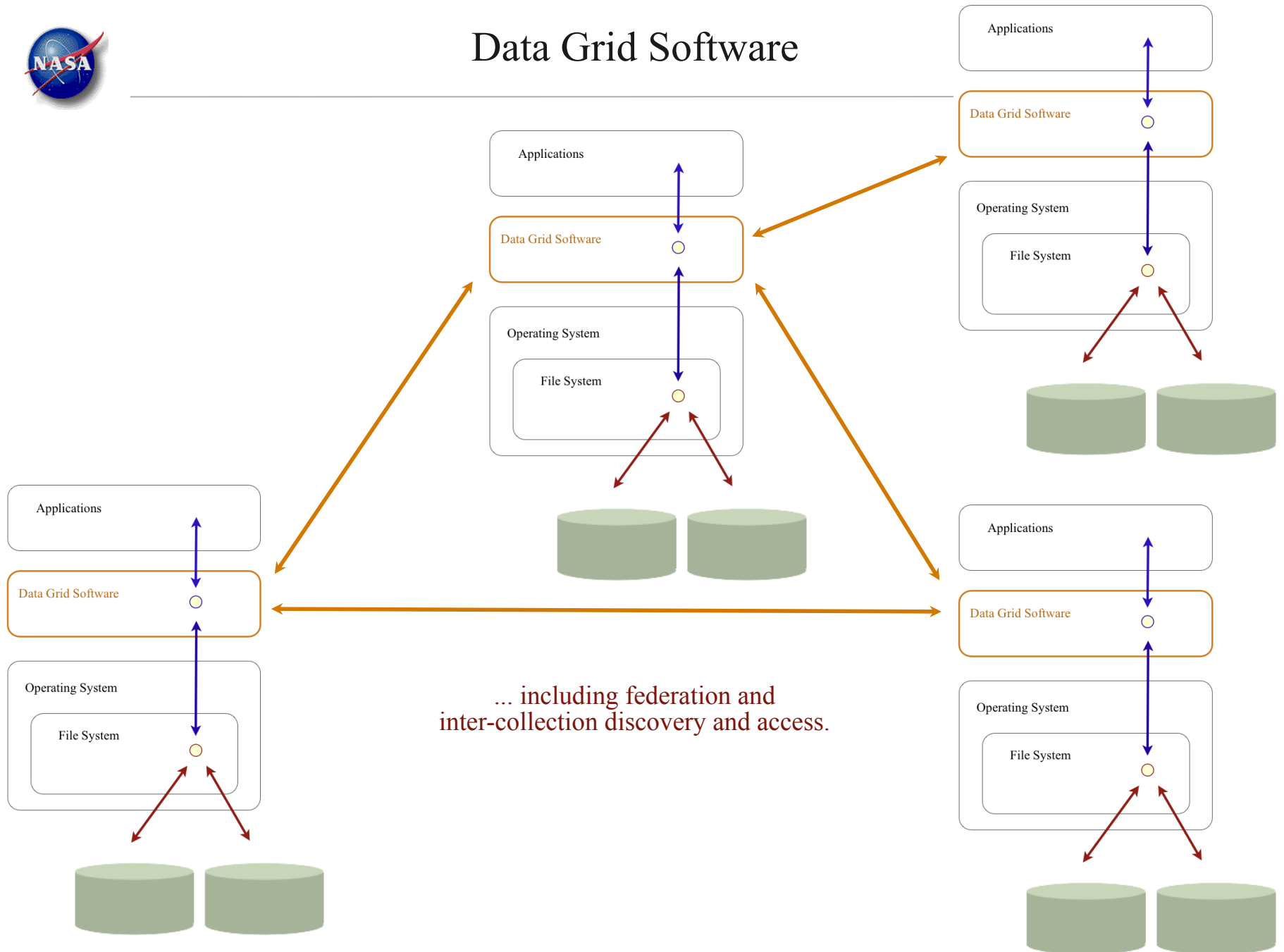


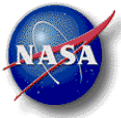
Name	Value	Unit
variables	Sea-level pressure, Surface pressure, Surface Geopotential, Geopotential height, Ozone Mixing Ratio	
title	MERRA reanalysis. GEOS-5.2.0	
source	Global Modeling and Assimilation Office. GEOSops_5_2_0	
references	http://gmao.gsfc.nasa.gov/research/merra/	
missing_value	9.9999999e+14f	
institution	Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, MD 20771	
history	File written by CFIO	
hdfEOSversion	HDFEOS_V2.14	
dimensions	TIME:EOSGRID = 1, YDim:EOSGRID = 144, XDim:EOSGRID = 288, Height:EOSGRID = 42	
conventions	CF-1.0	
contact	http://gmao.gsfc.nasa.gov/	
comment	GEOS-5.2.0	
checksum	91ddec7eee867abb8ca2e184ad2f8e92	

Data grid “middleware” runs as an application in user space and provides a richer set of metadata descriptors and extended capabilities ...



Data Grid Software





iRODS: integrated Rule-Oriented Data System

Background

- Open source data grid software system.
- Developed by the Data Intensive Cyber Environments (DICE) group, University of North Carolina.
- Historic roots in data grids, digital libraries, persistent archives, and real-time data systems R&D, and SRB.

Features

- Targets large repositories, large data objects, digital preservation, and integrated complex processing.
- Supports server-side workflows implemented by chaining execution rules together based on data policies.
- Enables scalability and extensibility.

Major Concepts

- Policies => iRODS rules.
- Mechanisms => iRODS microservices.

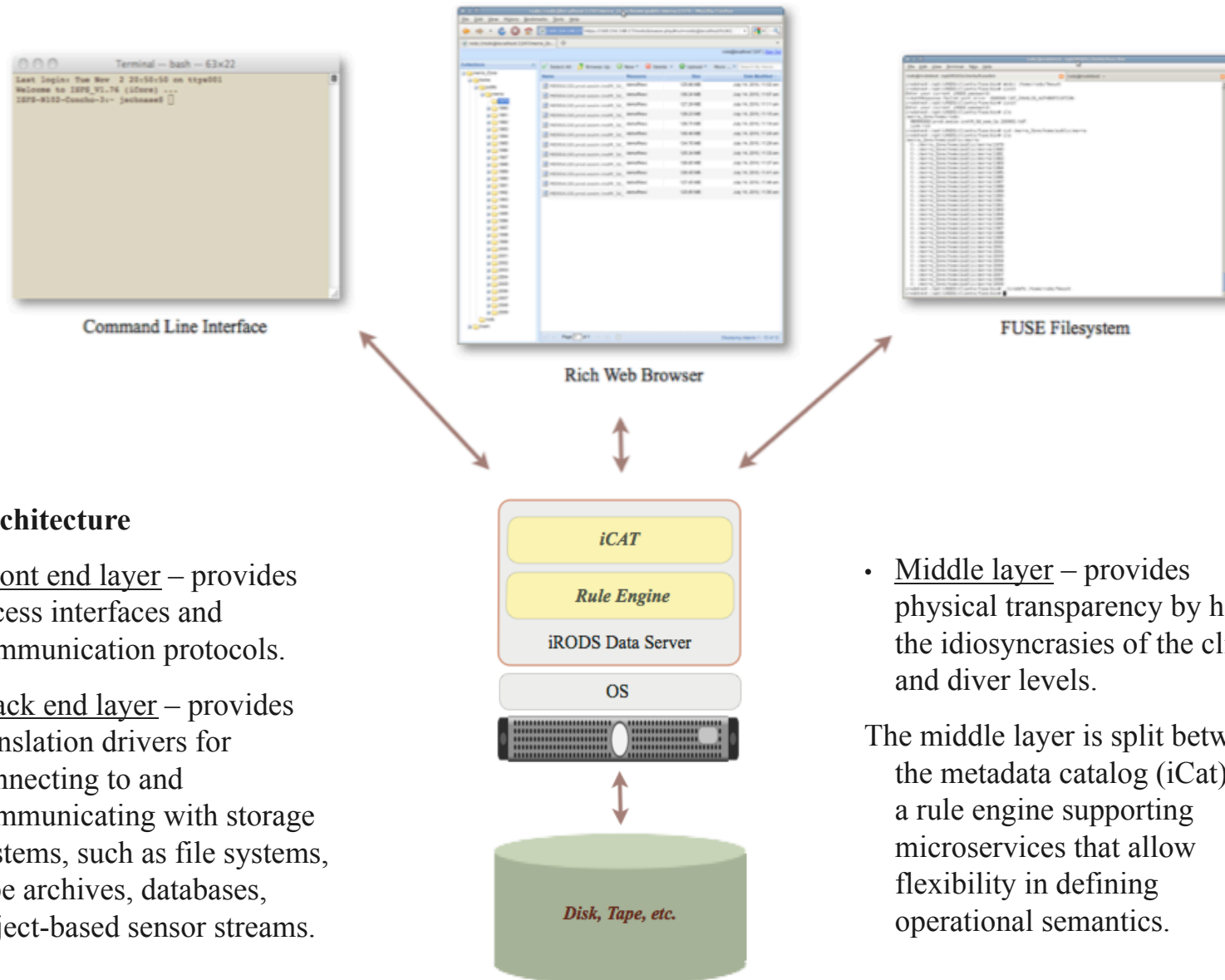
With iRODS metadata providing the information necessary to perform these mappings



www.irods.org

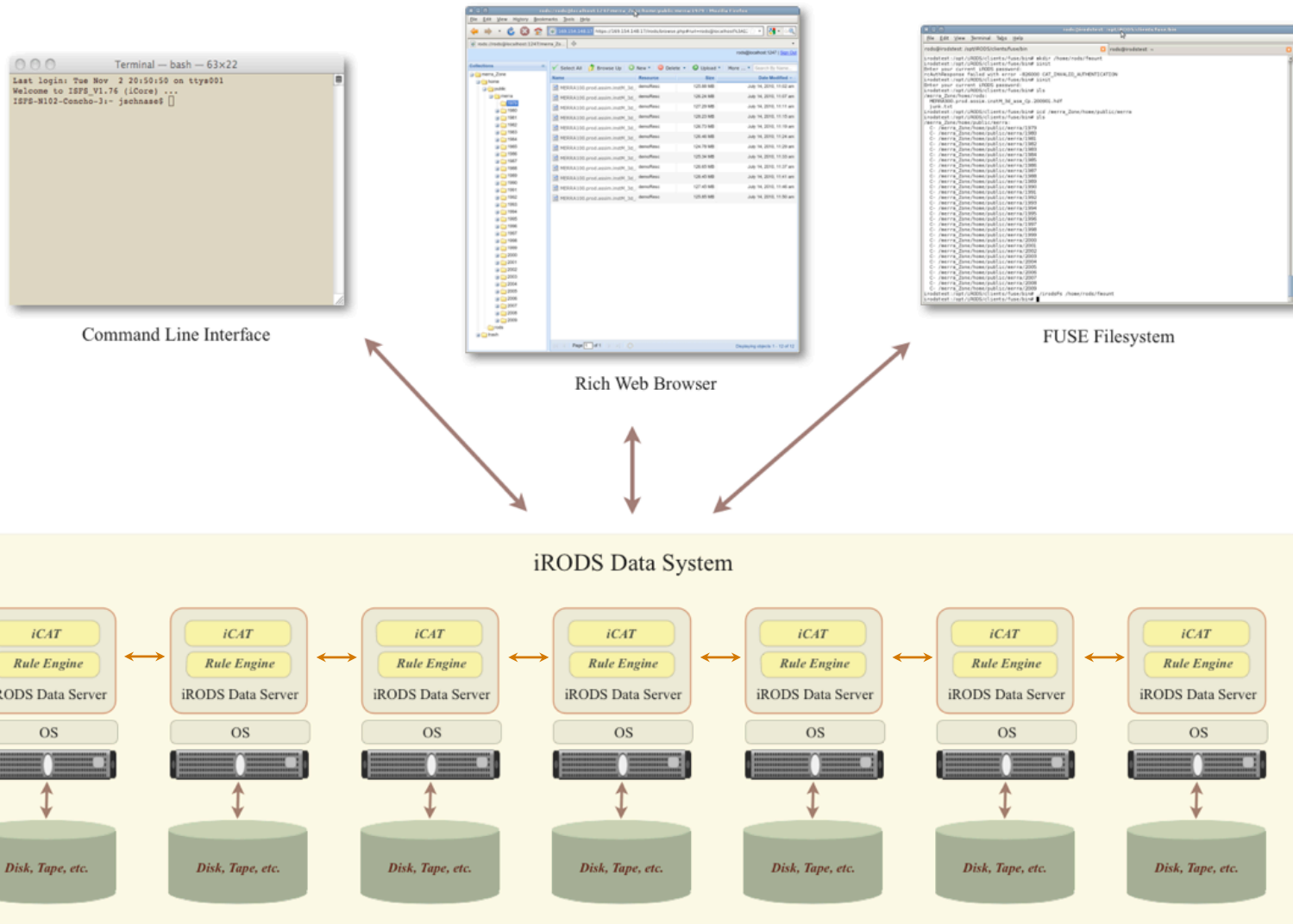


iRODS: integrated Rule-Oriented Data System



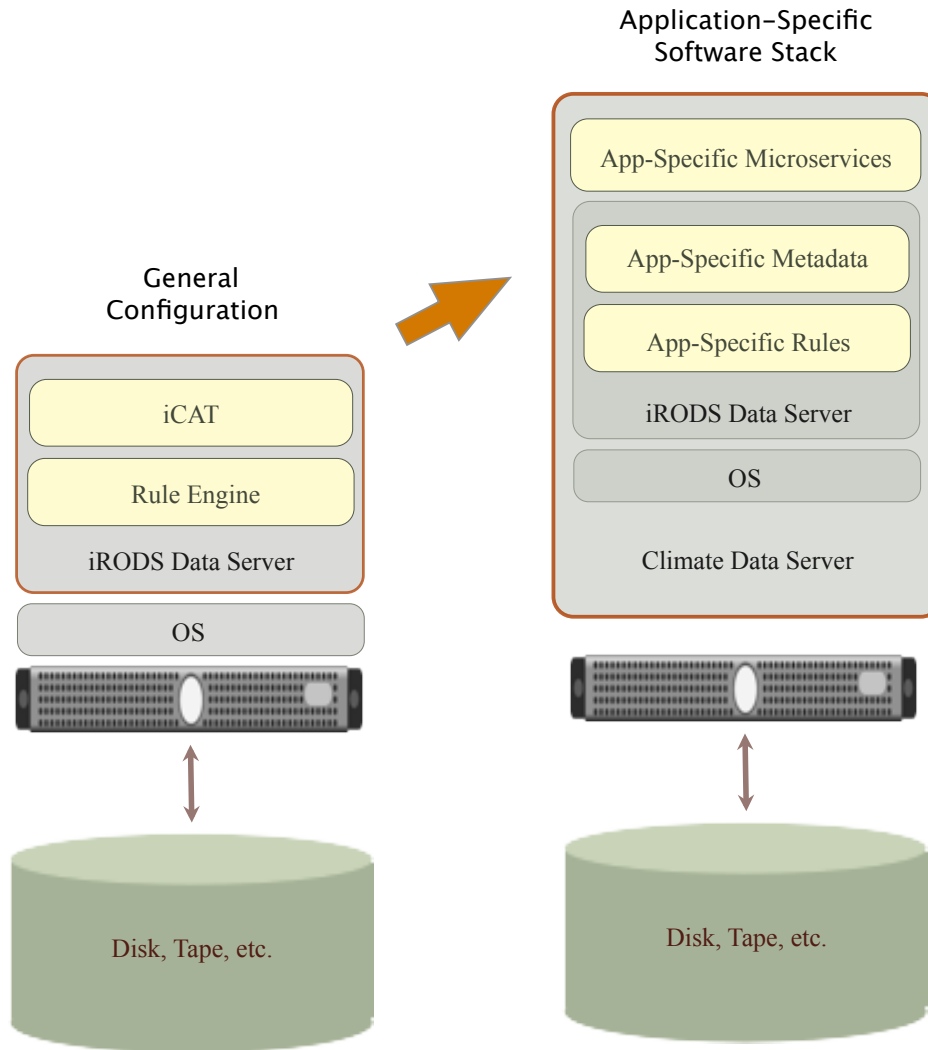


iRODS: integrated Rule-Oriented Data System





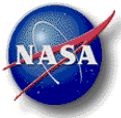
iRODS-Based Climate Data Server



Core Components

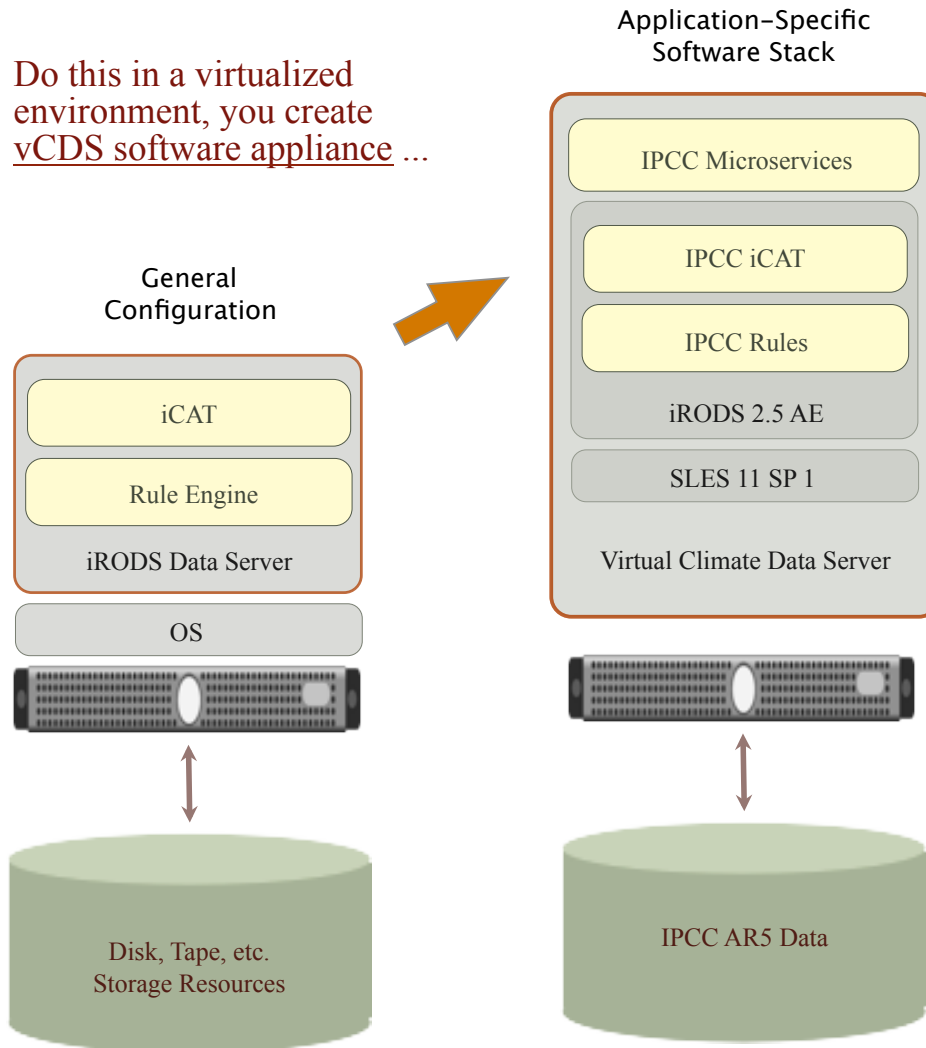
- Application-specific microservices
- Application-specific metadata
- Application-specific rules
- A specific release of iRODS
- A specific operating system

Do this in an organized way, you create a CDS software appliance ...



iRODS-Based Virtual Climate Data Server for IPCC Data

Do this in a virtualized environment, you create vCDS software appliance ...



vCDS 1.0 Product Suite

IPCC / NetCDF Module

- iRODS microservices, rules, configuration settings, and utilities required for canonical IPCC/NetCDF CRUD operations ...

Administrative Extensions

- iRODS Postgres extensions and utilities to log system-level object provenance and provide QA for OAIS metadata compliance ...

Repetitive Provisioning

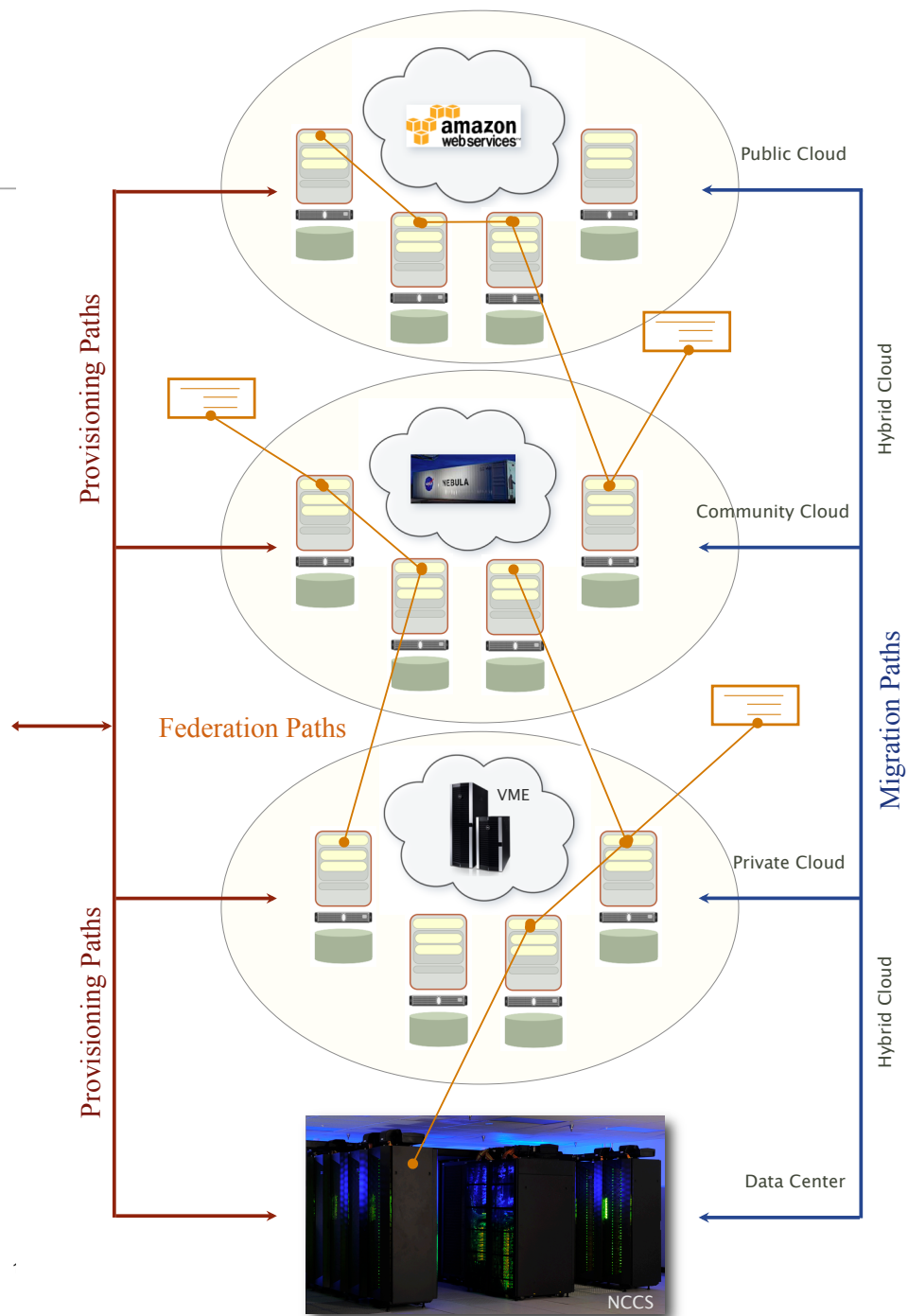
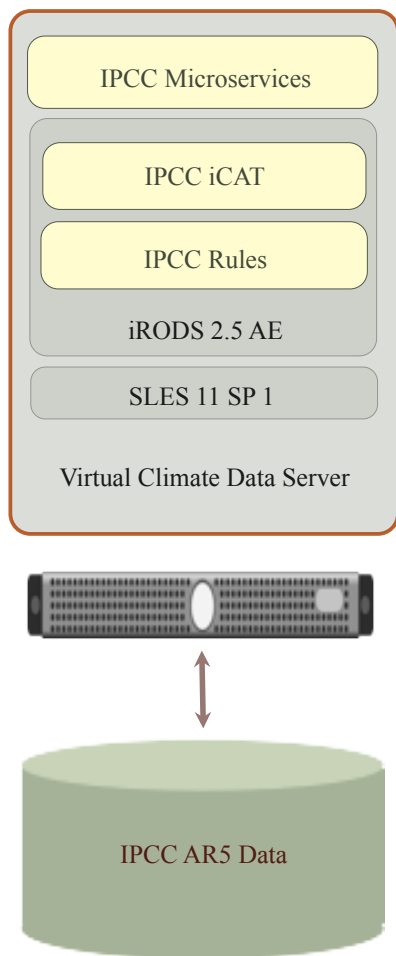
- RPM scripts to build software stacks for the SLES 11 SP1 (IaaS), iRODS AE (PaaS), and CDS/IPCC (SaaS) virtual images ...

Deployment and Distribution

- Product library, documentation, and SLA infrastructure for distribution, deployment, and help desk support ...



vCDS / VaaS Architecture





vCDS / VaaS Architecture

Virtualization is a driving concept

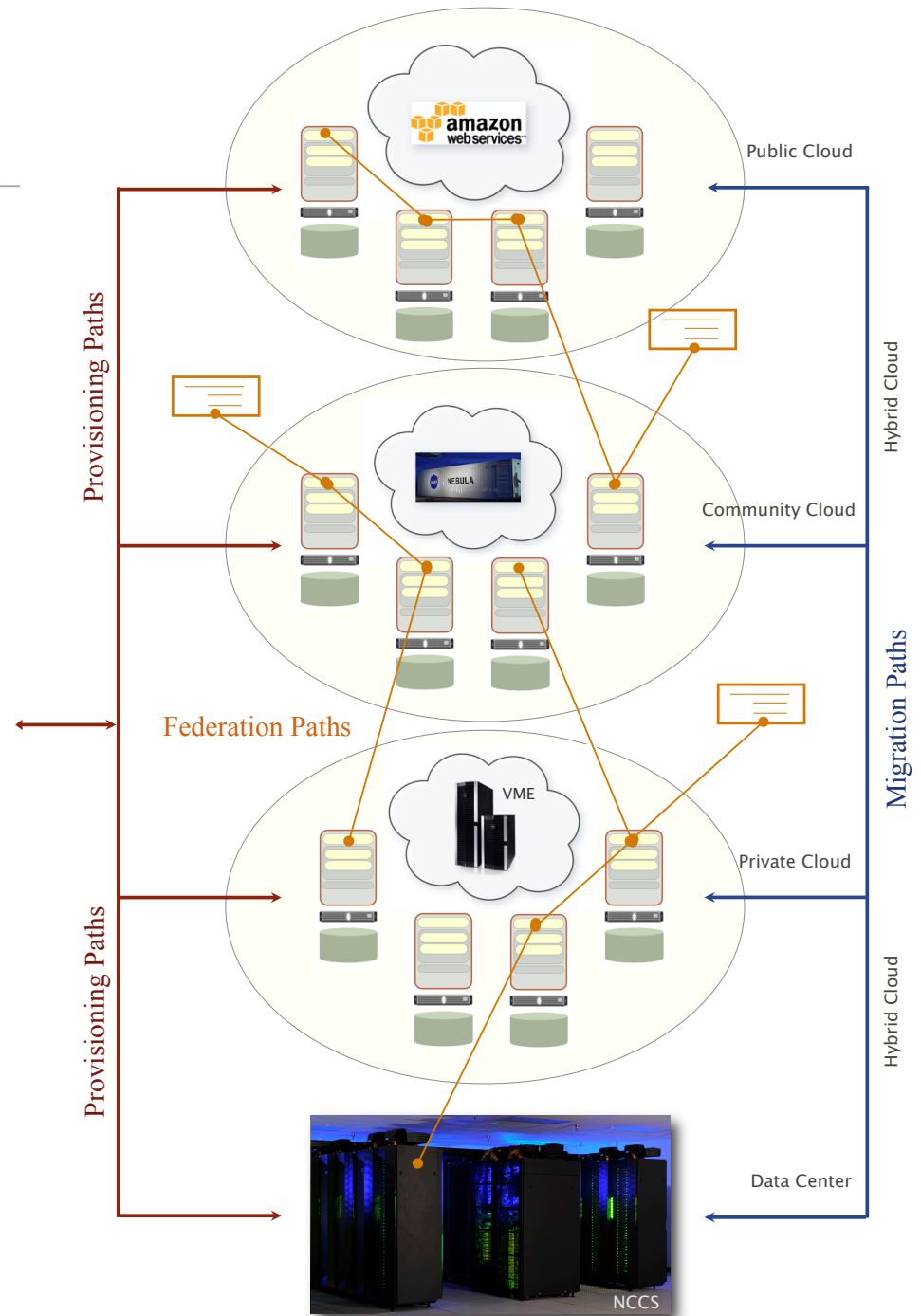
- Provides ready access to a tiered array of services that are flexible, adaptable, scalable, and stageable to NCCS “bricks and mortar” facilities as needed ...

Virtualization-as-a-Service is a huge unmet need in cloud computing

- This approach provides an agile entry point into the NCCS for new customers with data-centric requirements ...

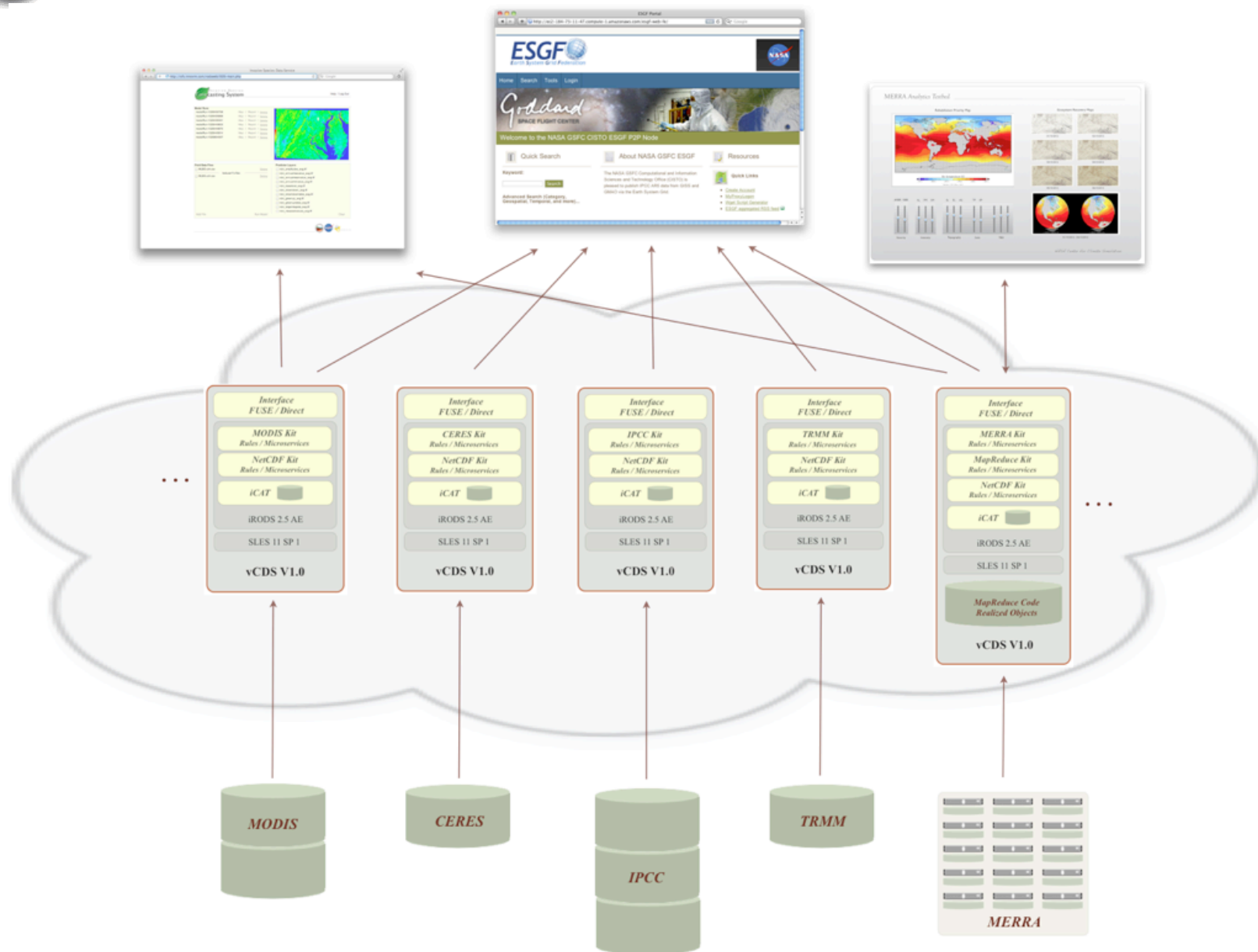
With VaaS, you can distribute or deploy

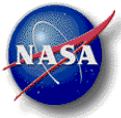
- If performance is an issue, you can do an old fashioned install. You can distribute images if you’re in a virtual world. You can also host images thereby enabling PaaS or SaaS ...





Ecosystem of Managed Collections





iRODS-Based Climate Data Services

John L. Schnase

*Office of Computational and Information
Science and Technology (Code 606)*

NASA Goddard Space Flight Center

September 12, 2012
